Journal of Educational Psychology
1989, Vol. 81, No. 3, 392-399

0022-0663/89/$00.75

The "Testing" Phenomenon: Not Gone but Nearly Forgotten

John A. Glover
Teachers College and Burris Laboratory School
Ball State University

The "testing" phenomenon refers to the finding that students who take a test on material between the time they first study and the time they take a final test remember more of the material than students who do not take an intervening test. 4 experiments examined the testing phenomenon in student's memory for brief passages and labels for parts of flowers. Experiments 1a and 1b demonstrated the generality of the phenomenon to the methods and materials used in the current study. Experiment 2 ruled out an "amount of processing" hypothesis as a way of accounting for the testing phenomenon. The results of Experiment 3 seemed to indicate that the testing phenomenon resided in the number of complete retrieval events. Experiments 4a, 4b, and 4c focused on the completeness of retrieval events and indicated that the influence of retrieval on later memory performance was determined, at least in part, by the completeness of the initial retrieval event.

Although there are numerous reasons for testing students' learning, one important factor is that tests improve students' memory for content (Modigliani & Hedges, 1987). In typical laboratory studies examining the so-called *testing phenomenon*, participants are given a list of words to learn. Then, after the initial learning episode, subjects are tested one or more times. The results of numerous studies consistently have demonstrated that subjects tested between the initial learning episode and the final test given over the material outperform subjects only given the final test (e.g., Cuddy & Jacoby, 1982; Jacoby, 1978; Runquist, 1986).

Surprisingly, given the number of laboratory studies that have examined the testing phenomenon in recent years, very little educationally relevant research has been done on the topic in quite some time. In fact, the best example of educationally relevant work in the area was reported 50 years ago in the *Journal*. Spitzer (1939) had all 3,605 sixth graders in nine Iowa cities read a passage and then tested them. The children were assigned to 1 of 10 conditions that differed in terms of how the initial learning episode, the intervening test(s), and the final test were arranged. Although Spitzer's outcomes were complex and the statistical analyses were less sophisticated than would be the case in contemporary research, a clear and general finding emerged: Intervening tests improved students' memory for the content.

The question of concern here is not so much whether tests enhance memory—the data overwhelmingly indicate they do. Instead, the emphasis is on why a test given between an initial learning episode and a final test enhances students' memory performance. A review of the literature suggests two major hypotheses: (a) amount of processing and (b) number of complete retrieval events. A brief description of these hypotheses follows.

I thank Ray Dean for critically reading an earlier version of this article.

Correspondence concerning this article should be addressed to John A. Glover, Teachers College and Burris Laboratory School, Ball State University, Muncie, Indiana 47306.

The simplest explanation for the effect of intervening tests is an *amount of processing* hypothesis (e.g., Kolers, 1973). In this view, memory performance is determined by the amount of processing devoted to specific bits of information. An intervening test merely causes students to process information for an additional time prior to a final test, thereby improving final test performance.

A second hypothesis, which may be referred to as the *number of complete retrieval events* hypothesis (or, simply, the *retrieval* hypothesis) holds that it is the number of complete retrieval events that influences final test memory performance. In this view, it is the processing engendered by acts of retrieval that accounts for the effects of intervening tests, not merely the amount of processing. Two predictions may be made on the basis of the retrieval hypothesis. First, two (or more) intervening tests may be more effective than one. Unlike the amount of processing hypothesis, however, this is believed to be true only when two tests are spaced apart so that they allow for complete retrieval operations in each instance. Second, retrieval operations requiring different levels of completeness (namely, free recall, cued recall, and recognition) should have varying effects on final memorability with the most complete retrieval operation (free recall) having the greatest influence.

Particularly germane to the retrieval hypothesis is the *spacing effect*. The spacing effect refers to the finding that recall for verbal information is better when learning trials are spaced rather than massed (Cuddy & Jacoby, 1982; Dellarosa & Bourne, 1985). Dellarosa and Bourne have conceived of the spacing effect in terms of an accessibility hypothesis. In their view, massed encoding trials do not require full encoding processes beyond the first encounter with new materials. This is because a record of the information gained in the first encounter remains accessible in memory. Consequently, subsequent encoding trials in massed sessions demand only that subjects review information already in memory. In contrast, full encoding processes are required for each trial when they are spaced apart. This is because forgetting or deactivation of information occurs between trials. Each time the material is

encoded in spaced trials, complete encoding processes are required.

Tests are not encoding trials. Yet the kind of reasoning used by Dellarosa and Bourne (1985) to account for differences between massed and spaced encoding trials also may be pertinent for a consideration of retrieval trials. That is, it may be that massed retrieval trials are less effective than spaced-apart retrieval trials because full retrieval processes are required only on the first retrieval trial in a massed session. On subsequent retrieval trials, the information necessary for the task is accessible in working memory, so subjects only need to review this information. In contrast, forgetting or deactivation should occur between spaced retrieval trials, thus necessitating full retrieval processes in each trial. At this point it is unclear whether retrieval is influenced in the same way as encoding. If so, it would seem that spaced-apart retrieval trials would be more beneficial to subsequent memory for the content than would massed retrieval trials.

The purpose of the current study was to examine the hypotheses described above in terms of how well they might account for the effects of intervening tests on students' memory. Overall, the focus was on students' memory for brief segments of prose and concepts being taught in class. Four experiments are presented as described below.

Experiments 1a and 1b were conducted to determine whether the testing phenomenon would be observed with the methods and materials used here. In Experiment 1a, undergraduates read and studied a brief passage. Those in one condition returned 2 days later for an intervening test. Those in another condition did not. Four days after the initial learning episode, all students were tested. Experiment 1b was similar to Experiment 1a except that it was conducted with middle-school students, used regular instructional materials, and varied the form of tests.

Experiment 2 was designed to test the amount of processing hypothesis. In this experiment, subjects in one condition encoded the material, received an intervening test immediately thereafter, and took a final test 4 days later. Subjects in a second condition performed the same tasks but took the intervening test 2 days after the encoding episode. Times for encoding and times for testing were held equal in the two conditions. Consequently, if the amount of processing hypothesis were correct, no difference would be observed on students' final test performance. If a significant difference between the conditions emerged, the amount of processing hypothesis could be ruled out.

Experiment 3 was based on the results of Experiment 2. Only the spaced retrieval condition in Experiment 2 resulted in greater levels of final test memory performance. One way to account for this outcome is to argue that it is the number of complete retrieval events that enhances subsequent memory performance and that complete retrieval only occurred during the spaced intervening test. When the intervening test immediately followed encoding, the subjects still had the appropriate information active and accessible for review. Therefore, they did not retrieve (or at least fully retrieve) the information from memory. Experiment 2, however, was only a partial test of the retrieval hypothesis. If it is the number of complete retrieval events that influences subsequent memory performance, then multiple retrievals would have a more powerful effect than single retrievals. This would be the case, however, only when multiple retrievals are spaced apart because of a spacing effect similar to that observed in encoding trials. In other words, full retrieval processes would occur only when enough time has elapsed between retrieval trials to allow deactivation or forgetting of the information used by subjects during retrieval. To test this hypothesis, then, students in Experiment 3 were assigned to one of four conditions: single intervening test, two massed intervening tests, two spaced intervening tests, or a control. The critical contrast was between the conditions in which students encountered two massed or two spaced-apart intervening tests. If students in the spaced-apart condition outperformed those in the massed condition, the retrieval hypothesis would be supported. If performance in these two conditions was similar, the amount of processing hypothesis would be supported.

Experiments 4a, 4b, and 4c were constructed to examine the second prediction of the retrieval hypothesis. This prediction was that different types of retrieval operations would have varying effects on final test performance with the most complete retrieval operations having the greatest effect. Furthermore, these varying effects would be seen regardless of the type of retrieval required at the time of final test. Four conditions were developed that varied the retrieval demands of the intervening tests. In Experiment 4a, the final test was free recall. In Experiment 4b, the final test was cued recall. In Experiment 4c, the final test was recognition. On the basis of previous work with word lists, it was predicted that each form of intervening test would enhance final test performance, no matter what form the final test took (Cuddy & Jacoby, 1982).

Experiments 1a and 1b

Experiments 1a and 1b were performed merely to establish the generality of the testing phenomenon. Two experiments were conducted so that both laboratory and applied settings would be used.

Method

Subjects and setting. In Experiment 1a, subjects were 30 undergraduates who participated for course credit. All testing took place in small groups of 3 to 8 students across a 2-week period. All data were gathered in a small conference room. In Experiment 1b, subjects were 27 seventh graders enrolled at a laboratory school who participated as a part of class activities. All testing took place in groups of 5 to 7 during a 1-week period. All data were gathered in the students' regular classroom and another classroom next to it.

Materials. The materials in Experiment 1a consisted of a brief, 300-word essay describing the fictitious nation of Mala. The six paragraphs originally were constructed by Bruning (1968), and each consisted of approximately 50 words with four sentences. The materials in Experiment 1b consisted of a drawing of a flower with 12 of its parts labeled. As a part of an upcoming life sciences unit, the students were to learn the parts of flowers.

Procedure. In Experiment 1a, subjects were randomly assigned to the control or experimental condition. Subjects in both conditions met with the experimenter for an initial encoding session and a final test session. Subjects in the experimental condition additionally met with the experimenter for an intervening test session. During the

initial encoding sessions, subjects received a brief set of written instructions and then the 300-word essay. All subjects were asked to read and study the essay carefully because they would be tested over its contents. Ten minutes were allowed for studying the essay. Subjects in the experimental condition received written instructions asking them to return in 2 days. Subjects in the control condition were asked to return in 4 days. When subjects in the experimental condition returned after 2 days, they were given a free-recall test over the essay. They were asked to do their best; partial recalls were encouraged. Ten minutes were allowed for this task, after which the subjects were dismissed and asked to return in 2 more days. Four days after the initial encoding, all subjects returned for a final test. This test also was free recall, and directions identical to those given for the intervening test were used.

In Experiment 1b the procedures were similar. In this instance, however, all students were conducted from their regular classroom to an adjacent room by the experimenter. There they were given a copy of the drawing of a flower with each of its parts labeled. The students were told that one of their upcoming activities would be to learn the parts of a flower and that these small group sessions could help them get a "head start." Then students were given 15 min to study the drawing. They were encouraged to draw their own version of the flower and to quiz themselves to see if the parts of the flower were in memory. After their studying was completed, all materials relevant to the experiment were collected. Two days after the initial encoding sessions, students in the experimental condition met with the experimenter and completed an intervening test. This test consisted of a version of the same drawing used during encoding with lines drawn to the various parts of the flower. In this instance, however, none of the parts were labeled, and students were asked to label the flower parts from memory. Ten minutes were allowed for this task, after which the materials were collected. All students completed a final test over the parts of the flower in their regular class 4 days after the initial encoding sessions. The final test was identical to the intervening test.

Results and Discussion

In Experiment 1a, the appearance of idea units from the essay (24 were possible) in subjects' final test protocols was scored by two raters ($\kappa = .93$). These data were then entered into a simple independent-samples t test, with conditions as the independent variable and idea units recalled as the dependent variable. The results, $t(28) = 6.65$, $p < .01$, $SE_M = .79$, indicated that subjects in the experimental condition ($M = 9.06$, $SD = 2.49$) recalled significantly more than subjects in the control condition ($M = 3.80$, $SD = 1.76$). In Experiment 1b, the number of labels students could provide for the parts of flowers was scored by two raters ($\kappa = 1.00$). These data also were entered into an independent-samples t test. The results, $t(25) = 2.77$, $p < .01$, $SE_M = .63$, indicated that students in the experimental condition recalled significantly more flower parts ($M = 4.45$, $SD = 1.55$) than students in the control condition ($M = 2.36$, $SD = 1.32$).

These results clearly indicated the presence of the testing phenomenon. The results, however, do not shed light on any of the hypotheses described earlier. To begin to examine these hypotheses, a second experiment was conducted.

Experiment 2

Experiment 2 was designed to test the amount-of-processing hypothesis. Although total processing time was held constant among conditions, the spacing of the initial encoding session, the intervening test, and the final test were varied. A significant difference among conditions would not support the amount of processing hypothesis and would instead favor the retrieval hypothesis.

Method

Subjects, setting, and materials. Subjects were 48 undergraduate volunteers participating for course credit. Data were gathered in small group sessions in a conference room. The materials were those used in Experiment 1.

Procedure. In general, the procedures were highly similar to those used in Experiment 1a. In this instance, however, subjects were randomly assigned to a control, a "massed," or a "spaced" condition. In the massed condition, subjects completed the initial encoding session and then immediately completed the intervening free-recall test. Four days later the subjects returned to complete the final test. In the spaced condition, subjects completed the initial encoding session, returned in 2 days for the intervening test, and then returned again after 2 more days for the final test. Subjects in the control condition had no intervening test. Ten minutes each were allowed for the initial encoding session in all conditions, the intervening test in both experimental conditions, and the final test in all conditions.

Results and Discussion

The final test protocols were scored as previously described ($\kappa = .92$). Then the data were entered into a one-way analysis of variance (ANOVA) with conditions as the independent variable and essay idea units recalled as the dependent variable. The results, $F(2, 45) = 55.05$, $p < .01$, $MS_e = 4.91$, indicated a significant difference among the conditions. A priori orthogonal t tests then were used to follow up this finding. The results, $t(45) = 11.42$, $p < .01$, revealed that subjects in the spaced condition recalled significantly more of the essay content on the final test than subjects in either the massed condition or the control condition. No other contrast was significant $t < 1$ (see Table 1).

Table 1
Results of Experiments 1, 2, and 3

Experiment/condition	*M*	*SD*
1a		
Control	3.80	1.76
Intervening test	9.06	2.49
1b		
Control	2.36	1.32
Intervening test	4.45	1.55
2		
Control	3.63	2.02
Massed test	3.44	1.77
Spaced test	10.69	2.57
3		
Control	2.67	1.49
One intervening test	4.50	1.92
Two massed tests	4.43	2.10
Two spaced tests	9.44	1.53

Note. Experiments 1a and 2 used idea units recalled from a brief essay as the dependent variable (24 were possible). Experiments 1b and 3 used the number of flower parts recalled as the dependent variable (12 were possible).

The results of Experiment 2 confirm those of Experiments 1a and 1b in terms of the appearance of the testing phenomenon. More important, however, the results seem to rule out the amount of processing hypothesis as an explanation for the testing phenomenon. When the intervening test was administered 2 days after the initial encoding session, memory performance on the final test was enhanced. In contrast, when the intervening test was administered immediately after the initial encoding session, no facilitation of final test performance was observed. This pattern of results clearly is inconsistent with the amount-of-processing hypothesis. Within our ability to control processing time, subjects in both conditions spent the same amount of time processing information. Our observations of students' test-taking behaviors in the massed and spaced conditions indicated no significant difference in the actual time spent on task of subjects in the massed (M = 9.88 min, SD = .54) and spaced (M = 9.92 min, SD = .49) conditions ($t < 1$).

One way to account for the pattern of results seen in Experiment 2 is to suggest that it was the act of retrieving information for the intervening test that influenced final test performance. In this view, the act of retrieval itself is beneficial to future memory performance. Furthermore, complete retrieval processes may be most beneficial to memory performance, whereas less complete retrieval efforts may be less beneficial. In Experiment 2, it seems likely that full retrieval processes occurred on the intervening test only when it was delayed for 2 days after the initial encoding session. When the intervening test immediately followed the initial encoding session, however, it seems probable that a less complete retrieval event was required. It seems reasonable to assume that in the massed condition at least some of the essay content was still in short-term memory at the onset of the intervening test. In particular, it seems likely that the gist of the essay was available as well as the last few ideas mentioned in the text (Just & Carpenter, 1987).

If a retrieval hypothesis does account for the testing phenomenon, then two kinds of predictions can be made. First, multiple intervening tests may be more effective than single intervening tests. This should be true, however, only when multiple intervening tests are spaced apart so that they allow for complete retrieval operations in each instance. Second, different types of retrieval operations (free recall, cued recall, and recognition) may have varying effects on final test performance. The most complete retrieval operation (free recall) should have the greatest effect on final test performance. Experiment 3 was conducted in order to examine the first of these predictions and Experiment 4, to examine the second.

Experiment 3

Experiment 3 was designed to determine whether two spaced-apart intervening tests were more effective than either a single intervening test or two massed intervening tests. If two spaced-apart intervening tests led to greater levels of recall than two massed intervening tests, the results would support the retrieval hypothesis.

Method

Subjects, setting, and materials. Subjects were 57 seventh-grade students attending a consolidated, rural middle school. The students participated as a part of their normal science course. All data were collected in the students' regular classrooms and a room adjoining it. The materials were those used in Experiment 1b.

Procedure. Students were randomly assigned to one of four conditions: single intervening test, two massed intervening tests, two spaced intervening tests, and control. All students were assigned the flower drawing as an in-class activity and given 15 min to study and redraw it in preparation for a future test (the parts of flowers were to be a major part of the students' next unit, to begin approximately 8 days after the experiment). After the study time, the materials were collected. Students in the single intervening test met with the experimenter the next day in small groups of 5 or 6. Students left class to come to these meetings in an adjoining room. During these meetings, students were given a blank drawing of a flower and were asked to label as many parts as they could from memory. Ten minutes were allowed for this activity, after which the students returned to class. Students in the two massed intervening tests condition generally encountered the same procedures except that after the first intervening test they were given exactly the same test a second time. They were told that this procedure "might help them keep from forgetting the parts of flowers by providing practice in remembering." Ten minutes were allowed for each of the two tests. The students in the two spaced intervening tests condition completed the first intervening test 1 day after the initial encoding session. After completing the task, they were dismissed. They returned the next day for the second intervening test, which consisted of exactly the same task. The experimenter recorded instances of student off-task behavior during all testing sessions. Students in the control condition received no intervening test. All students received their final test in class on the 5th day of the experiment. The final test was the same flower-labeling task.

Results and Discussion

The number of flower parts correctly labeled by students was scored by two independent raters (κ = 1.00), and the data were entered into a one-way ANOVA, with conditions as the independent variable and flower parts correctly labeled as the dependent variable. The results, $F(3, 53) = 38.16$, $p < .01$, $MS_e = 3.16$, indicated a significant difference among the conditions (see Table 1). The orthogonal t tests indicated that students in the two spaced intervening tests condition recalled significantly more of the flower parts on the final test than students in any of the other conditions $t(55) = 12.08$, $p < .01$. No significant difference was observed between students in the single intervening test condition and students in the two massed intervening tests condition, $t < 1$. Students in each of these conditions, however, recalled significantly more flower parts than students in the control condition, $t(55) = 3.27$, $p < .01$.

The results of Experiment 3 clearly confirm the prediction made concerning the effects of two intervening tests. That is, two intervening tests were significantly more beneficial to final memory performance than one intervening test, but only when the two intervening tests were spaced apart. When two intervening tests were massed, final memory performance was not improved beyond what was achieved by means of a single

intervening test. This result was observed even though equal times were allocated for each test, and no observable differences in on-task behaviors were seen among students. Using Dellarosa and Bourne's explanation (1985) for the spacing effect in encoding as a heuristic, a retrieval effects explanation for the results of Experiment 3 seems fairly clear. Retrieval processes facilitate subsequent retrieval. Each additional retrieval process performed on a set of materials further enhances the memorability of the material. In Experiment 3, full (or complete) retrieval processes were required on two occasions only when the intervening tests were spaced apart. When two intervening tests occurred one after the other, only one instance of full retrieval processes was required. Consequently, performance was not better than when only one intervening test was given. This is because the retrieval demands of one intervening test and two massed intervening tests were highly similar.

Although the results of Experiment 3 confirmed one of the predictions made on the basis of the retrieval effects hypothesis, another prediction remained to be tested. For that reason, a fourth experiment was conducted.

Experiments 4a, 4b, and 4c

A second prediction that can be made on the basis of a retrieval effects hypothesis is that the most complete retrieval operations should have the greatest influence on final test performance. Furthermore, this influence should be observed regardless of the type of retrieval operation required on the final test. Experiments 4a, 4b, and 4c were designed to examine this prediction.

Although it has been argued that recall and recognition involve different processes (see Klatsky, 1984, for a review), a dual-process model such as Anderson's (e.g., 1985) distinguishes among free recall, cued recall, and recognition primarily on the basis of the number of points at which memory searches may begin and the way in which activation spreads. From such a perspective, the amount or completeness of processing required by retrieval tasks increases from recognition to cued recall to free recall. Simply, individuals are presented many more points from which to enter memory in recognition tasks than in recall tasks. Consequently, free-recall tasks require fuller, more complete retrieval processing than cued-recall or recognition tasks. The results of Experiments 2 and 3 point to the possibility that the completeness of retrieval influences the subsequent memorability of the information—memory was facilitated only when intervening tests engendered complete retrieval processes. If the influence of intervening tests on final memory performance is governed by the completeness of the retrieval processes, then it seems reasonable to argue that free recall should result in the greatest benefit to later memory performance followed by cued recall and then recognition. Furthermore, this pattern of results should be observed regardless of whether the final test is a free-recall, cued-recall, or recognition test. On the other hand, if the results indicated that final test performance was greatest when the type of retrieval on the intervening test and the final test matched (as per Tulving's reasoning, 1985), the conclusion would be that final test performance was most influenced by a match of processing operations and that intervening tests influenced subjects' responses to final tests by means of some form of processing-context phenomenon.

Method

Subjects, setting, and materials. Each of the three experiments used 65 undergraduate volunteers who participated for course credit. As a result of missed sessions or incorrectly followed directions, the three experiments included 59, 62, and 64 subjects, respectively. Subjects participated in small groups of 5 to 8 and met in two small conference rooms for all their sessions. The stimulus materials were those used in Experiment 1a. The intervening cued-recall test consisted of 12 sentences taken from the essay (the first and last sentences from each paragraph) that were paraphrased and ended with a blank (e.g., "Mala's form of government is _____."). The intervening recognition test consisted of 12 sentences, 6 sentences drawn from the essay (the 2nd sentence in each paragraph) and 6 false sentences, all 12 in random order. The final cued-recall test used in Experiment 4b consisted of 24 paraphrased sentences, each ending with a blank, taken from the essay. The final recognition test used in Experiment 4c consisted of 48 sentences presented in random order: the 24 sentences contained in the essay and 24 distractors taken from other Mala material.

Procedure. The three experiments were run concurrently. Of the 195 students recruited for participation, 65 were randomly assigned to each of the experiments. In each experiment, subjects again underwent random assignment, this time to one of four conditions: free-recall intervening test, cued-recall intervening test, recognition intervening test, and control. As in Experiment 2, the order of events was the initial encoding session followed by the intervening test 2 days later in the experimental conditions. In all conditions, the final test occurred 4 days after encoding. The experiments varied in that Experiment 4a used a free-recall final test. Experiment 4b used a cued-recall final test, and Experiment 4c used a recognition final test.

Results and Discussion

The results from the three experiments were analyzed separately. In each instance the data were analyzed in an ANOVA, with conditions as the independent variable. In Experiment 4a the dependent variable was idea unit recall, in Experiment 4b the dependent variable was cued-recall scores, and in Experiment 4c the dependent variable was the number of correctly recognized sentences. Table 2 summarizes the data from Experiments 4a, 4b, and 4c.

In Experiment 4a, the analysis revealed a significant difference among conditions, $F(3, 55) = 25.64$, $p < .01$, $MS_e = 5.48$; see Table 1). The orthogonal t tests indicated that subjects in the free-recall condition recalled significantly more idea units on the final test than subjects in any of the other conditions, $t(55) = 11.69$, $p < .01$. In addition, subjects in the cued-recall condition recalled significantly more on the final test than the subjects in the control condition, $t(55) = 4.10$. Furthermore, there was no significant difference between subjects in the cued-recall and subjects in the recognition conditions, $t(55) = 1.06$.

When the results from Experiment 4b were considered, the analysis indicated a significant difference among conditions,

Table 2
Results of Experiment 4

Experiment/condition	*M*	*SD*
4a		
Control	3.71	1.99
Free recall	10.75	2.79
Cued recall	8.44	2.26
Recognition	7.20	1.90
4b		
Control	2.93	1.81
Free recall	7.31	2.23
Cued recall	5.31	2.02
Recognition	4.07	2.08
4c		
Control	6.07	2.52
Free recall	11.53	3.99
Cued recall	8.93	2.57
Recognition	8.07	2.46

Note. Experiment 4a used idea units recalled from a brief essay as the dependent variable (24 were possible). Experiment 4b used cued recall (24 were possible), whereas Experiment 4c used recognition (24 were possible).

$F(3, 58) = 11.89$, $p < .01$, $MS_e = 4.46$ (see Table 1). The results of orthogonal contrasts indicated that subjects in the free-recall condition had significantly higher final test cued-recall scores than subjects in any other condition, $t(58) = 8.43$, $p < .01$. In addition, subjects in the cued-recall condition had significantly greater final test cued-recall scores than subjects in the control condition, $t(58) = 4.02$, $p < .01$. No other contrasts reached significance.

The data from Experiment 4c similarly were entered into a one-way ANOVA. The results, $F(3, 60) = 7.85$, $p < .01$, $MS_e = 9.43$, indicated a significant difference among the conditions (see Table 1). Orthogonal contrasts indicated that subjects in the recall condition recognized significantly more of the chapter's content than subjects in any other condition, $t(60) = 9.32$, $p < .01$. In addition, subjects in the cued-recall condition recognized significantly more essay content than subjects in the control condition, $t(60) = 3.08$, $p < .01$. No other contrast was significant.

In each instance, the pattern of memory performance was as follows: The free recall intervening test resulted in greater levels of memory performance than the test, which led to greater performance than the recognition intervening test, which led to memory performance superior to the control condition. Even though the difference between the cued-recall intervening test condition and the recognition intervening test condition was not significant, the pattern of results would seem to indicate that it was the completeness of the retrieval process engaged in during the intervening test that influenced final test performance.

It also should be noted that consideration was given to treating Experiments 4a, 4b, and 4c as separate segments of a more complex Experiment 4. In such a combined approach, "proportion remembered" would have served as the dependent variable. This, however, would have introduced the unavoidable bias of treating free recall, cued recall, and recognition as though scores on these measures were directly comparable.

General Discussion

The results of the current study demonstrate the robustness of the testing phenomenon and begin to suggest theoretical accounts for its appearance. Experiments 1a and 1b showed the generalizability of the testing phenomenon to the methods and materials used in the current study. Experiment 2 determined that an intervening free-recall test enhanced subsequent memory performance, but only when the intervening test was spaced apart from the initial encoding session. Because processing time was carefully controlled in both the spaced and massed conditions, an amount-of-processing hypothesis was rejected as an explanation for the testing phenomenon. In Experiment 3, two intervening tests had a more facilitative effect on final memory performance than a single intervening test but only when the two intervening tests were spaced apart. This result suggests that it was the retrieval process and, more specifically, the number of complete retrieval events that influenced subsequent memory performance. The results of Experiments 4a, 4b, and 4c all indicated that free-recall intervening tests had a significantly more facilitative influence on subsequent memory performance than cued-recall or recognition intervening tests. Cued-recall intervening tests and recognition intervening tests both had significantly beneficial effects on final test performance. Finally, this general pattern was observed on free-recall, cued-recall, and recognition final tests. Like the results of Experiment 3, the results of Experiment 4a, 4b, and 4c seem to support the retrieval hypothesis.

It should be noted that there is no absolute way of dealing with processing time. Even though the same amount of time was allocated to each testing session and even though no observable differences in processing (or, more correctly, on-task behavior) time turned up, it is still possible that subjects in different conditions used their time in different ways. Such difficulties in completely accounting for subjects' processing times, however, seem unavoidable. Still, because consistent results appeared over each of the replications, the current results do indeed seem to rule out processing-time hypotheses.

It also is important to note a potential bias pointed out by one of the reviewers of this manuscript. That is, in Experiment 4, where differences on final test performance could have been due to the fact that the cued-recall and the recognition intervening tests required that only a portion of the materials be remembered, whereas subjects who received the free-recall intervening test ostensibly could have remembered the entire passage. This bias would make the separation of an amount of processing from a number-of-complete-retrievals hypothesis very difficult, because attenuated demands from cued-recall or recognition intervening tests would engender lower amounts of processing than free-recall intervening tests. This bias presumably could be addressed by broadening the cued-recall and recognition intervening tests. At some point, however, the sheer amount of time involved in various forms of intervening tests would seem to result in an additional confound.

Unlike early work in educational settings (Spitzer, 1939), the focus of the current study was on a theoretical accounting for the testing phenomenon. The current study also confirmed some of the findings of more recent laboratory research,

particularly the appearance of the spacing effect in retrieval trials (Modigliani & Hedges, 1987). Unlike recent laboratory research, however, the current work focused on to-be-remembered materials that are much closer to typical classroom assignments (i.e., brief essays and labels for the parts of a flower).

The results of the current study suggest that it is the number and completeness of retrieval events that influence subsequent memory performance. Two converging arguments seem to support this conclusion. First, two intervening tests were more facilitative than one intervening test only when the intervening tests were spaced apart. If Dellarosa and Bourne's (1985) reasoning about encoding trials can be extended to retrieval trials, it can be argued that only when the two retrieval trials were spaced apart did full retrieval occur in both instances. When the two intervening tests were massed, in contrast, full retrieval occurred only during the first intervening test. The second intervening test, which immediately followed the first, required only a shallow reprocessing of content in working memory.

The second argument supporting the idea that the number and completeness of retrieval events influence subsequent memory performance is based on the effects of different kinds of intervening tests on final test performance. Presumably, free recall requires a more complete retrieval episode than cued recall, which itself requires a more complete retrieval episode than recognition (see Anderson, 1985). The results of Experiments 4a, 4b, and 4c indicated that free recall had a more facilitative influence on subsequent retrieval performance than cued recall and that cued recall had a larger, although not a significantly larger, influence on memory performance than recognition. This pattern was observed when free recall, cued recall, and recognition were used on the final test. These results suggest that the completeness of intervening retrieval events influences subsequent retrieval, with the more complete intervening retrieval events having a more beneficial effect than less complete intervening retrieval events. In addition, the results cast doubt on potential contextual explanations for the testing phenomenon.

Overall, then, the results of the experiments reported here suggest it is the number of retrieval events and their completeness that set the parameters for the testing phenomenon. In general, it seems reasonable to hypothesize that, as the number of retrieval events between initial encoding and final testing increase, so too should final memory performance. This hypothesis should only hold, however, for complete retrieval events. Furthermore, it also seems reasonable to expect that the law of diminishing returns must set in at some point. That is, after some as yet undetermined number of intervening retrievals, subsequent memory performance probably will not be further enhanced. In addition, it also can be hypothesized that complete intervening retrieval events (i.e., free recall) will have a more beneficial effect on subsequent memory performance than less complete (i.e., recognition) intervening retrieval events. This hypothesis, however, must be tempered with consideration of the relative difficulty of various intervening retrieval events (e.g., the possibility that some recognition tasks will be more difficult than recall tasks, and, therefore, will require more complete retrieval).

How does the retrieval event actually influence memory? Although there is no clearly acceptable explanation for what is involved in retrieval, the findings of a post hoc analysis may be helpful. That is, an analysis of subjects' protocols indicated that the material recalled on the final tests always was a subset of the materials recalled on the intervening tests. That is, in the experimental conditions, nothing was recalled on the final tests that had not been recalled on the intervening tests. This prompted a speculation that was based on a dual-process model of retrieval (e.g., Anderson, 1985). That is, it can be postulated that intervening tests have their effect on subsequent retrieval because the first retrieval event "unitizes" the set of items retrieved. The nodes activated fully enough to allow for retrieval have the activation pathways among them strengthened and, perhaps, increased. On subsequent retrievals, then, the set of items retrieved previously have closer links to one another and the retrieval of any one of this set increases the likelihood of retrieving any other of the set. Furthermore, such unitization only includes previously recalled items. This speculation accounts for the items on subsequent retrieval trials always being a subset of an initial retrieval. Those items not retrieved on the first retrieval are not part of the "unit" activated during later retrievals and so are not readily available in memory.

It also should be noted that retrieval need not be seen as having unique or unusual effects on recall to account for the results of the current study. It is possible that retrieval simply is a special form of rehearsal (see Modigliani & Hedges, 1987, for an argument favoring this position). It may be that retrieval events merely allow for an additional full processing of the material (see Dellarosa & Bourne, 1985) much as spaced encoding events do. Furthermore, this perspective could account for the fact that the items remembered during the final test always are a subset of the items remembered on the intervening test. Simply stated, the items retrieved on the first intervening test become the set of items rehearsed for subsequent tests. Those items not retrieved on the first intervening test are not likely to be retrieved in the future because they did not receive additional rehearsal. The results of Experiments 4a, 4b, and 4c do not fit neatly into this perspective because no match-of-operations effect was observed as would be expected if retrieval events were merely special cases of rehearsal. Still, this explanation cannot be ruled out completely.

The two speculations described above could be contrasted, it seems, by examining the role of various prompts of different strengths on subsequent memory for the material and by using various probe techniques more sensitive than recall tests. In any event, regardless of whether retrievals are special instances of rehearsal or unique cognitive activities, the results clearly indicate that intervening tests enhance student's memory for content and that it is not the sheer amount of rehearsal involved that brings about the improved memory.

References

Anderson, J. R. (1985). *The architecture of cognition.* Cambridge, MA: Harvard University Press.

Bruning, R. H. (1968). Effects of review and testlike events within

the learning of prose material. *Journal of Educational Psychology, 59*, 16–19.

Cuddy, L. J., & Jacoby, L. L. (1982). When forgetting helps memory: An analysis of repetition effects. *Journal of Verbal Learning and Verbal Behavior, 21*, 451–467.

Dellarosa, D., & Bourne, L. E. (1985). Surface form and the spacing effect. *Memory and Cognition, 13*, 529–537.

Jacoby, L. L. (1978). On interpreting the effects of repetition: Solving a problem versus remembering a solution. *Journal of Verbal Learning and Verbal Behavior, 17*, 649–667.

Just, M. A., & Carpenter, P. A. (1987). *The psychology of reading and language comprehension*. Boston: Allyn & Bacon.

Klatsky, R. L. (1984). *Human memory: Structures and processes* (2nd ed.). San Francisco: Freeman.

Kolers, P. A. (1973). Remembering operations. *Memory and Cognition, 1*, 347–355.

Modigliani, V., & Hedges, D. G. (1987). Distributed rehearsals and the primary effect in single-trial free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13*, 426–436.

Runquist, W. N. (1986). The effect of testing on the forgetting of related and unrelated associates. *Canadian Journal of Psychology, 40*, 65–76.

Spitzer, H. F. (1939). Studies in retention. *Journal of Educational Psychology, 30*, 641–656.

Tulving, E. (1985). On the classification problem in learning and memory. In L. Nilsson & T. Archer (Eds.), *Perspectives on learning and memory* (pp. 73–101). Hillsdale, NJ: Erlbaum.

Received January 6, 1988
Revision received February 2, 1989
Accepted March 22, 1989 ■